ビジネス応用におけるデータマイニング志向情報システム

Data Mining Oriented Information Systems for Business Application

矢 田 勝 俊
Katutoshi YADA

近年、ビジネスの世界において、社内で蓄積されている大量のデータから有益な知識を発見しようとする要求は、次第に大きくなってきている。しかし従来の研究では、開発したアルゴリズムを現実のビジネスのデータに適用し、実際に有益な知識を発見している例は少ない。我々は過去の企業との共同研究において、大規模なデータベースからの有益な知識発見に関する研究を継続的に行なってきた。Hamuro他（1998）やYada他（1998）、Etoh他（1997）の中では、顧客の購買行動に基づいた様々な有益な知識を発見するシステム、ならびに関連の要素技術について明らかにしている。これらは、現実のビジネス世界での大量のデータ処理に関する問題解決のプロセスで生まれてきたアイデアが、上記の知識発見を可能にしたことを示している。本稿の目的は、これらの有益な知識発見を可能にした基礎となるシステム設計思想、（一般的なコンセプト）を提示することである。

これまでのKDDやデータマイニングの研究は、与えられたデータセットの中でいかに効率よく、よいルールを見つけるかという問題に焦点を当ててきた。知識発見の視点から、そのようなデータセットが、どのように蓄積されるべきかを議論したものは、見当たらない。我々は、データマイニングやKDDといった発見科学の研究を現実に適用する立場に立って、それを可能にするための、システム設計思想、具体的にはデータの蓄積・獲得の方法について、議論する。

実際に企業で採用されている設計思想に基づいたシステムでは、入力されているデータの多くが消失している。なぜなら、それらのシステムでは事前に業務に必要な項目（属性）を決め、入力されたデータをその項目ごとに整形・蓄積し、他のデータは不必要なデータとして消失されてしまうからである。こうした考え方は、知識発見を現実に行うにあたって、重要な問題を引き起こすことになる。つまり、知識発見に必要なデータは事前に予測できるわけではないために、実際の分析プロセスで必要なデータが入手できないという問題である。

そこで我々は、データマイニングのような将来の詳細な分析に答えるために、すべての業務操作・プロセスを記録する、データ構造、ならびにその設計思想全体として、「履歴ベース」を提示する。従来のシステムと履歴ベースの関係は、スナップショットとビデオの記録の関係に似ている。ある人物の笑顔を記録するという目的において、スナップショットは彼の笑顔を記録することができるだろう。しかし、その写真からその笑顔の理由を推測することは難しい。たとえ、笑顔より前の状況を移したスナップショットであっても、その写真と写真の間に、何が起こったかを推測することは、その写真からは不可能

である。それに対して、ビデオは笑顔が生じる一連の出来事すべてを記録するものである。したがって、ある程度その理由を推測することが可能になる。その基礎となる考え方は、現象を解決するには、インプットやアウトプットだけを把握するのでなく、すべてのプロセスを把握することが必要であるというものである。その出来事のプロセスが「履歴」であり、履歴ベースとは、「履歴」を重視し、それらを蓄積・検索できるデータ構造、システム設計思想全体を我々は「履歴ベース」と呼んでいる。

この履歴ベースは、２つの履歴、「操作履歴」と「思考履歴」から構成されている。操作履歴とは、コンピュータに入力されたデータや操作を含む一連の出来事のプロセスである。例えば、POSレジにおける販売行為、データ分析における検索行為などが含まれる。思考履歴とは、コンピュータには入力されてない、一連の出来事のプロセスである。例えば、人の購買動機や、検索の結果出てきたデータの分析者の評価が含まれる。こうした出来事に関する詳細な履歴が蓄積されることによって、通常、多様なデータ要求が発生する知識発見に対して、我々は対応し、知識発見に成功してきた。データマイニングやKDDといった発見科学の研究は、その材料であるデータを無視しては、現実への適用は困難である。この論文は、データマイニングやKDDといった発見科学を現実に適用するという観点から必要とされる、データ構造や蓄積方法といったデータ環境に関わる基礎的な技術を提示することを目的としている。

我々はこれまで、企業と共同で大規模なデータベースから有益な知識を発見するための研究を行ってきた中で、以下の例のような重要な問題を認識することができた。それは、与えられたデータからいかにして有益なルール（知識）を発見するのかという問題と同じくらいに、分析に必要なデータをどのように蓄積し、また新たに必要になったデータをどのように獲得していくかという問題が重要であるという事実である。

なお、本研究の成果の一部を論文誌“Discovery Science（1998）PP.441-442”に公表している。